DOCUMENT RESUME

ED 060 058 24 TM 001 144

AUTHOR Monk, Janice J.
TITLE Another Look at the Relationship Between Frequency of Testing and Learning.
INSTITUTION Illinois Univ., Urbana. Office of Instructional Resources.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
REPORT NO RR-307
PUB DATE Oct 69
CONTRACT OEC-3-7-0611740-0271
NOTE 12p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *College Students; *Learning; Physical Geography; Quality Control; *Relationship; Research Criteria; *Research Design; Research Problems; Statistical Analysis; *Testing; Test Results

ABSTRACT

Several experiments investigating the relationship between frequency of testing and learning are reviewed, and features of design which could have influenced outcomes are discussed. A study on test frequency is presented, in which relevant variables, apparently not considered in previous research, are controlled. Results indicate that moderate variations in test frequency do not significantly affect learning. (MS)

ED 060058

research report # 308

research report

Another Look at the Relationship Between Frequency of Testing and Learning

by

Janice J. Monk

Measurement and Research Division
507 East Daniel Street, Champaign
University of Illinois

October, 1969

U.S. DEPARTMENT OF HEALTH, EDUCATION, & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Research Report #307

Another Look at the Relationship Between Frequency of Testing and Learning[1]

Janice J. Monk

1. The research and evaluation reported herein was performed pursuant to a contract (OEC-3-7-0611740-0271) with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of P.L. 83-531, cooperative research.

The relationship between frequency of testing and learning by college students has been studied by a number of investigators. Their research has shown that varying test frequency may have no effect, a slight effect, or a significant effect on learning (Balach, 1964). From a review of some of these experiments, it appears that this variation in results may be related to the nature of the experimental design, and not to the relationship between test frequency and learning per se.

The present study will review several experiments and discuss features of design which could have influenced outcomes. It will then report on a study of test frequency which attempts to control relevant variables apparently not considered in the previous research.

Review of Research

Only two of the studies reviewed indicated unreservedly that more frequent testing during a semester increased student learning. Pikunas and Mazzota (1965) compared the effects of weekly testing with no testing over two six-week periods in chemistry classes of high school seniors. Identical exams, given at the end of each period, were used as criteria for measuring the effects of the experimental treatment. Results showed higher scores for the group receiving weekly tests although the statistical significance of the differences was not reported.

One feature of the experimental design casts doubt on whether Pikunas and Mazzotta (1965) were measuring the effects of test frequency. Each group had met for three 68 minute periods per week. Apparently, the experimental group had one lecture period, one test period, and a third period devoted

to grading and reviewing tests. The control group had one lecture, but spent the remaining time in private study, in recitation (concerning homework) and in a question-and-answer period. Thus it is questionable whether test frequency alone was the variable being manipulated.

The second study (Fitch, Drucker, & Norton, 1951), claiming significant differences, did attempt to control for two other confounding variables. Fitch et al. (1951) partialed out differences between students in the experimental and control groups with respect to (a) attendance at optional discussion sections and (b) a measure of student ability and interest in the subject matter (government). Both of these variables might have influenced scores on the criterion measure (performance on monthly tests). However, when these variables were controlled for, a significant difference remained between the experimental group (which had weekly amd monthly tests) and the control group which had monthly tests only. In the Fitch et al. (1951) study, differences in the class activities of the experimental and control groups were negligible -- except for the ten minutes per week devoted to testing the experimental group.

There were other features of the design which make it questionable whether test frequency alone was being considered. One problem in this, and all other experiments reviewed, is the interpretation of the concept of test frequency. Most investigators seem to be studying the effects of the amount or quantity of testing rather than of fequency of testing. Thus in the Fitch el al. (1951) study, not only was there a difference in the frequency of testing but there were also differences (however small) in the time allocated to testing and in the nature and extent of review.

Three other studies showed either slight or no significant differences between experimental and control groups. In these studies the preceding comments about test frequency remain applicable. There were, moreover, other confounding variables.

In two of the studies, Eurich et al. (1937) and Standlee and Popham (1960), a large proportion of the items from pre-test and mid-term examinations was repeated on the criterion measures (mid-term and final examinations). Further, in the Eurich et al. (1937) study, weekly quiz items (given to the experimental group) were also included in the criterion.

Yet another confounding variable to consider is the style of test items used in the experimental control and criterion measures. Standlee and Popham (1960) used twenty true-false items for each weekly quiz, but multiple choice items in mid-term and final examinations. Selakovich (1962) used twelve "pop" quizzes (presumably of the objective type) on his experimental group and gave no quizzes to his control group. Three teacher-made essay examinations and a standardized test served as criteria. Obviously, there needs to be some consideration given to the effects of repeating items and of using tests with different item styles as criteria before the effects of test frequency can be isolated.

The applicability of the results from some of these experiments to the classroom situation is doubtful. Few teachers are likely to choose between twelve "pop" quizzes or no class quizzes as alternative models. Likewise, the situation where two of the three class meetings are devoted to test related activities is unusual.

Method

The subjects for this investigation were 164 students in an introductory Physical Geography course at the University of Illinois. Students attended two lectures per week in one of two classes taught by the same instructor. In addition, they attended one of eight quiz-discussion sections which each met for one hour three days per week. These quiz-discussion sections were taught by four graduate teaching assistants.

In designing the experiment it was assumed that quiz sections could be regarded as equivalent with respect to the students' ability and prior knowledge of the subject matter. That is, the assignment of subjects to treatments was considered random. There is supporting evidence for this assumption. Registration in particular classes was mainly a function of timetables, and students know nothing in advance about assignments of teaching personnel nor testing procedures. Few, if any, students enter the course with prior college geography courses and most have had little background in the material covered. Examination of records from a number of semesters has shown no significant differences between quiz sections on pretest or posttest scores.

The possible effects of teacher variance were controlled by designating one of the two classes taught by each assistant as an experimental group and the other as a control.

Since the goal of this study was to vary test frequency, and not the amount of testing nor the class time devoted to testing and other activities, it was decided to give an equal number of test items to each group and to cover the same material on the tests for both groups. The control sections received eight fifteen-item quizzes and the experimental sections were

given four thirty-item quizzes. In this study the designation of control and experimental groups was arbitrary. Because the accustomed procedure in geography was the giving of weekly quizzes, those groups who received eight quizzes were termed control groups. Two hour-long tests of 80 items each and a comprehensive 200 item final examination were the criteria. Three equivalent (at least in terms of content) versions (A, B, and C) of each criterion measure were developed to administer in the two lecture sections.

The pattern of administration of quizzes was such that the control group was tested at approximately weekly intervals -- three times before the first "hourly", three times before the second "hourly", and twice before the final examination. The experimental group received quizzes at approximately monthly intervals -- one before the first hour exam, two before the second and once before the final. Both of these patterns were considered appropriate models for a normal classroom situation.

It was assumed that, under the testing conditions established, the control group would be exposed to frequent testing and review. The experimental group, however, still would have some opportunity to experience the type of test items and material to be used in the criterion measures, but would not have frequent class testing and review.

All tests were constructed by the experimenter in an attempt to ensure consistency of style, coverage, and difficulty between experimental and control groups. Item statistics, for many items, were available from previous test administrations. Questions were true-false, multiple choice and short answer, with a similar balance of item types for both groups. Each quiz contained factual and problem solving items in about a two to

one ratio. Criterion tests were similar in style to class quizzes. Quizzes were returned to students after one week. One-third of each student's final grade depended on quiz scores.

Results

Mean raw scores on each of the hour tests and final examinations were compared for students from the experimental and control sections. Mean differences were tested, using t-tests with adjustments for heteroegeneous variance (following Edwards, 1963) where necessary. Table 1 shows the results of this analysis.

Insert Table 1 about here

In only one case was there a significant difference at the .05 level.

Discussion

The results of this experiment suggest that moderate variations (weekly versus monthly) in test frequency do not significantly affect learning, as measured by criteria which are similar to class quizzes. Only one significant difference ($p < .05$) was found between an experimental group and its control group. This may well have been a chance result since nine separate t tests were calculated. An alternative explanation is supported by that one significant difference and the larger mean scores of all three control groups on the first "hourly" examination criterion. For the first hour long examination, the differences in class treatments -- one quiz versus

three -- might have been great enough to influence criterion scores. This effect was noted by Standlee and Popham (1960); in their study also, it disappeared by the end of the semester. It may be that (a) the student does not need many cues to discern the type of learning required by a course and (b) that, if his learning is being measured by "hourly" and final examinations, frequent quizzes are of little importance to his learning once he knows what will be demanded of him.

No known characteristic of the various experimental and control groups was seen by the investigator as a source of heterogeneous variance in two of the t test comparisons.

At this stage there seems little to be gained from further experimentation of the preceding type. However, there are a number of other interesting questions related to test frequency. How do students feel about their own needs for frequent testing? To date, most studies of test frequency have involved learning factual material. How are other types of learning related to frequency of testing and review? Perhaps test items (and answers) might, more appropriately, be regularly included within the context of the learning material, rather than only at the end of units of study. A recent paper by Bruning (1968), following similar research by Rothkopf (1966), suggests the value of incorporating test-like events within prose materials as an aid to learning. Finally, one might suggest that regular testing could be used by the teacher, not only as a stimulus to student study and as a measure of student learning, but also for diagnosing student difficulties with learning materials and for assessing the effectiveness of his own teaching.

References

Balach, J. "The Influence of the Evaluating Instrument on Students' Learning," American Education Research Journal, 1964, 1, 169-182.

Bruning, R. H. "Effects of Review and Test-Like Events Within the Learning of Prose Materials," Journal of Educational Psychology, 1968, 59, 16-19.

Edwards, A. L. Statistical Methods for the Behavioral Sciences, Holt, Rinehart and Winston, New York, 1963.

Eurich, A. C., Longstaff, H. P., and Wilder, M. "The Effects of Weekly Tests upon Achievement in Psychology" in The Effective General College Curriculum as Revealed by Examinations. A report of the Committee on Education Research, University of Minnesota, University of Minnesota Press, Minneapolis, 1937.

Fitch, M., Drucker, A. J., and Norton J. A. Jr. "Frequent Testing as a Motivating Factor in Large Lecture Classes", Journal of Educational Psychology, 1951, 42, 1-20.

Pikunas J. and Mazzota, D. "The Effects of Weekly Testing in the Teaching of Science." Science Education, 1965, 49, 373-376.

References

(cont.)

Rothkopf, E. Z. "Learning from Written Instructive Material: An Explanation of the Control of Inspection Behavior by Test-Like Events," American Educational Research Journal, 1966, 3, 241-249.

Selakovich, D. "An Experiment Attempting to Determine the Effectiveness of Frequent Testing as an Aid to Learning in Beginning College Courses in American Government." The Journal of Educational Research, 1962, 55, 178-180.

Standlee, L. S. and Popham, W. J., "Quizzes Contribution to Learning", Journal of Educational Psychology, 1960, 322-325.

Table 1

Comparisons between Control and Experimental Groups on Two Hour-Long Tests and a Final Examination

Criteria	Comparison Groups[1]	N	$\bar{X}$	S.D.	t	p
First Hour Examination	A Control	24	51.16	8.66	.8[2]	> .05
	A Experimental	36	40.55	18.95		
	B Control	29	52.65	6.73	2.13	< .05
	B Experimental	25	48.76	6.78		
	C Control	27	47.29	9.56	.05[2]	> .05
	C Experimental	23	46.60	6.56		
Second Hour Examination	A Control	25	45.24	9.60	.002	> .05
	A Experimental	39	45.20	10.00		
	B Control	29	46.44	15.33	-.86	> .05
	B Experimental	26	49.30	10.59		
	C Control	24	50.62	10.69	1.08	> .05
	C Experimental	21	47.47	9.67		
Final Examination	A Control	23	123.17	19.40	.84	> .05
	A Experimental	43	118.70	22.40		
	B Control	25	130.60	23.23	.38	> .05
	B Experimental	23	128.00	23.35		
	C Control	30	130.10	24.14	-.32	> .05
	C Experimental	21	132.00	17.44		

1. Letters A, B, and C refer to equivalent (at least in terms of content) forms of the same test.

2. Because of heterogeneous variance, the value of t required for significance was found following Edwards (1963).